Detection of change points in underlying earthquake rates, with application to global mega-earthquakes

Detection of change points in underlying earthquake rates, with application to global mega-earthquakes

Sarah Touati[1], Mark Naylor[1], Ian Main[1]

[1] School of GeoSciences, University of Edinburgh, Grant Institute, James Hutton Road, Kings Buildings, Edinburgh EH9 3FE

Accepted date. Received date; in original form date

Abbreviated title: Has the global rate of mega-earthquakes changed

Corresponding author:
Sarah Touati,
School of GeoSciences, University of Edinburgh, Grant Institute, James Hutton Road, Kings Buildings, Edinburgh EH9 3FE
sarah.touati@ed.ac.uk
+447812242882

Summary

The recent spate of mega-earthquakes since 2004 has led to speculation of an underlying change in the global 'background' rate of large events. At a regional scale, detecting changes in background rate is also an important practical problem for operational forecasting and risk calculation, for example due to volcanic processes, seismicity induced by fluid injection or withdrawal, or due to redistribution of Coulomb stress after natural large events. Here we examine the general problem of detecting changes in background rate in earthquake catalogues with and without correlated events, for the first time using the Bayes factor as a discriminant for models of varying complexity. First we use synthetic Poisson (purely random) and Epidemic-Type Aftershock Sequence (ETAS) models (which also allow for earthquake triggering) to test the effectiveness of many standard methods of addressing this question. These fall into two classes: those that evaluate the relative likelihood of different models, for example using Information Criteria or the Bayes Factor; and those that evaluate the probability of the observations (including extreme events or clusters of events) under a single null hypothesis, for example by applying the Kolmogorov-Smirnov and 'runs' tests, and a variety of Z-score tests. The results demonstrate that the effectiveness among these tests varies widely. Information Criteria worked at least as well as the more computationally-expensive Bayes factor method, and the Kolmogorov-Smirnov and runs tests proved to be the relatively ineffective in reliably detecting a change point. We then apply the methods tested to events at different thresholds above magnitude $M \geq 7$ in the global earthquake catalogue since 1918, after first declustering the catalogue. This is most effectively done by removing likely correlated events using a much lower magnitude threshold ($M \geq 5$), where triggering is much more obvious. We find no strong evidence that the background rate of large events worldwide has increased in recent years.

Keywords

Statistical seismology; seismicity and tectonics; earthquake interaction, forecasting and prediction; probabilistic forecasting

1. Introduction

The ability to detect changes in the basic rate of earthquakes, which can be caused by crustal fluid movement, volcanism, human activities, or as yet unknown mechanisms, is an important yet challenging part of statistical analysis of earthquake occurrence. Examples include distinguishing accelerating trends from background processes in volcanic eruption (e.g. Bell *et al.*, 2013); changes in rate due to re-injection of waste water produced by developing unconventional hydrocarbon reserves (e.g. Ellsworth, 2013); and localised changes in rate due to Coulomb stress redistribution following large natural events (e.g. Hainzl *et al.*, 2010). Recently there has been considerable practical interest in assessing the significance of apparent rate changes, generated by the recent cluster of mega-earthquakes, beginning in 2004, generating some debate as to whether this is a real global change (Ammon *et al*., 2010) or a statistical artefact of sampling a stationary distribution (Michael, 2011). A variety of statistical methods are routinely used either to claim significance in rate changes in local (Lombardi *et al.*, 2006, Hainzl and Ogata, 2005) or global (Bufe and Perkins, 2005) earthquake occurrence, or to demonstrate that apparent changes are nevertheless consistent with a temporally stationary stochastic process (Touati *et al.*, 2014, Michael, 2011).

One of the reasons that event rate variations are difficult to infer is that the background process must first be separated from the aftershocks. To this end, the Epidemic-Type Aftershock Sequences (ETAS) model (Ogata, 1988) is often used in statistical analysis of earthquakes. It is based on a Poisson process of 'independent' events, representing the effect of stationary tectonic loading, with aftershocks triggered from every event in the catalogue (including aftershocks of other events). Aftershocks occur at a time-decaying rate defined by the Omori law; the total number of aftershocks increases exponentially with the magnitude of the triggering event. The ETAS model is thus capable of accounting for event rates that fluctuate in time and space due to aftershock triggering, according to its conditional intensity function:

$$\lambda(t|H_t) = \mu + A \sum_{i:t<t_i} \exp(\alpha(M_i - M_0))(1 + \frac{t - t_i}{c})^{-p} \quad (1)$$

where t_i are the times of the past events and M_i are their magnitudes, μ is the independent or 'background' event rate, A, c and p are parameters of the Omori law, and α is the aftershock productivity parameter. Aftershocks themselves are thus a form of clustering or non-stationarity, which presents the challenge of accounting both for the aftershock process (under sometimes considerable uncertainty) and any additional non-stationarity caused by a change in the seeding rate, μ, as two separate effects. We have previously shown that the degree of temporal overlap of separate aftershock sequences is an important determinant of the accuracy with which the aftershock statistics can be inferred using maximum likelihood (Touati *et al.*, 2011); overlap results in a masking effect whereby the signature of the aftershocks in the time intervals between events is lost. Thus it can be challenging to account for aftershocks accurately enough to be confident of event rate changes that go beyond ordinary aftershock triggering.

There are further, more subtle pitfalls associated with changepoint detection: one being that an extreme event (of some kind) in a very long data set is actually an expected outcome in a random process. Thus what appears to be an anomaly can sometimes turn out to have a non-negligible probability of occurring within the period spanned by the data set. Furthermore, as Shearer and Stark (2012) point out, every realisation of a random process contains some feature or other that is statistically highly unlikely, and so the more specifically an 'anomalous' feature of a data set is described, the more unlikely it appears to be. We should be sceptical about claims involving very narrow descriptions of highly anomalous occurrences that are only drawn up after having seen the data.

In this paper we test the efficacy of various well-referenced statistical methods for detecting change points or choosing between stationary or non-stationary models for seismicity. In principle these methods could be applied to the detection of changes in background rate in natural earthquake populations, volcanic seismicity or induced seismicity. While the latter two examples represent clear physical perturbations to the long-term stationary tectonic forcing observed in Nature (DeMets, 1995),

it is also possible to have 'mode-switching' between two metastable stationary points in the emergent dynamics of systems with stationary driving forces, including statistical physics models of faulting and earthquakes. For example Cowie et al. (1993) show clear mode switching of deformation on different fault systems at different times in their model for fault systems, based on a resistor network analogue model for normal fault evolution by repeated earthquakes.

This spatial element of the problem introduces complexities, not least in how one defines an anomalous region objectively, and without prejudicing the outcome with conscious or unconscious selection bias. In this paper the focus is on a comprehensive analysis of the methods of rate change detection in a case where no spatial data selection is used. At this stage of model development, and given the relatively short temporal sample of the process (Naylor et al., 2009) we limit our interest to models with a single change point in time. This provides a simple control test on the methods themselves, but we acknowledge the need to extend the analysis to regional data and more complex temporal models in due course. The statistical methods for determining non-stationary behaviour methods largely fall into two types: those that evaluate the likelihood of different models (which involves direct evaluation of probability densities at particular points), and those that evaluate the probability of the observations, or something more extreme, under a null hypothesis (which involve, instead, the area under part of a distribution curve). We test the methods on simulations of Poisson processes and the ETAS model, using both stationary simulations and simulations with changepoints. The main novelty in the paper is to apply a fully Bayesian approach to the multi-dimensional inversion problem in earthquake rate statistics, using the Bayesian 'evidence' term as a discriminant between models with different numbers of parameters, made possible by advances in affordable computational power. We then compare the performance of the resulting Bayes factor to other model selection criteria from the Bayesian or frequentist staples, including proxies such as the Akaike and Bayesian Information criteria. We then go on to use to the insights gleaned in applying the more successful methods to the question of whether the rate of megaquakes worldwide has increased in a way that cannot be accounted for with a stationary process which includes correlations introduced by well-understood triggering processes.

Section 2 presents the statistical methods evaluated in this study, while section 3 explains the testing methodology employed to discover how well these methods perform on synthetic data. Section 4 presents the results of the testing. In Section 5 we address the question around global megaquakes, describe the data used, and present results of the statistical methods. Section 6 is a general discussion of outstanding issues and recommendations.

2. Changepoint detection methods

The changepoint detection methods tested in this study fall into two types: those that evaluate the likelihood of different models, and those that evaluate the probability of the observations (or something more extreme) under a null hypothesis, i.e. a single model. In the first category, Information Criteria and the Bayes Factor are used to assess the relative likelihood of competing hypotheses. In the second category, we look at the Kolmogorov-Smirnov and runs tests, which are general tests for deviations from a Poisson process, and a variety of Z-score tests, which are more specific to the situation of a changepoint and test the significance of deviations from a long-term Poisson process. We also look at the basic Poisson probability of a number of events in a time window given a long-term event rate, showing that a simple Z statistic can give equivalent results to this.

AIC and BIC

The Akaike and Bayesian Information Criteria (AIC and BIC, respectively) are both metrics that represent the relative goodness of fit of different models, based on the maximised log-likelihood of each model. Both metrics penalise the use of more complex models to inhibit over-fitting.

AIC (Akaike, 1974) is given by:

$$\mathrm{AIC} = -2 \times \max(\mathrm{LL}) + 2k \tag{2}$$

where LL denotes the log-likelihood, and k is the number of parameters in the model. A lower AIC value denotes a better fit to the data of one model over another. In the case that one of the parameters

is a changepoint, the contribution to the penalty depends on how the changepoint was chosen: if it is pre-set with some *a priori* knowledge, there is no penalty for it, whereas if it is inferred solely from the data, its penalty depends on the size of the data set, asymptotically approaching the value 3 (Ogata, 1992). The use of AIC in the selection of stationary or non-stationary models for seismicity is widespread (Llenos and Michael, 2013, Ma, 2001, Hainzl and Ogata, 2005, Ogata, 2007, Matsu'ura and Karakama, 2005, Katsumata, 2011, Zhuang *et al.*, 2005, Ogata, 1999, Ogata, 1998, Zhuang, 2000).

BIC (Schwarz, 1978) depends more explicitly on the size of the data through the number N of data points:

$$\mathrm{BIC} = -2 \times \max(\mathrm{LL}) + k \ln N \tag{3}$$

Its penalty tends to be stricter than that of AIC, so that simpler models tend to be preferred more often.

The difference in AIC or BIC between models is more meaningful than the particular AIC or BIC values, and it is the difference that is used as a model choice criterion. A rule of thumb for significance of the AIC difference is that it should be at least 4 (Burnham & Anderson, 2002).

Bayes Factor

A more fully Bayesian approach than the information criteria above involves the use the Bayesian evidence or marginal likelihood (Sambridge *et al.*, 2006). This is the denominator, *p(d)*, in Bayes' theorem:

$$p(x|d) = \frac{p(d|x)p(x)}{p(d)} \tag{4}$$

where 'd' refers to data and 'x' to model parameters, so that $p(d|x)$ is the likelihood, $p(x)$ the prior on the parameters, and $p(x|d)$ the posterior probability of the parameters given the data. The quantity $p(d)$ is often regarded merely as a normalising constant, more fully expressed as:

$$p(d) = \int p(d|x)p(x)dx \tag{5}$$

However, it is also a measure of how good a fit the model is to the data, as it represents the likelihood over the entire model space averaged with respect to the prior. Competing models can be compared by computing the Bayes Factor, which is the ratio of evidence values between the two models (Kass and Raftery, 1995), the model with higher evidence being preferred. This is somewhat analogous to AIC and BIC, which are based on the difference of the log likelihoods; for the evidence, the likelihood is integrated over the prior rather than maximised. It is in some ways more satisfactory: the issue of uncertainty in the maximised likelihood is removed; and there is no need for the inclusion of a (heuristic) penalty to enforce parsimony. Occam's razor is inherent in the evidence, because although a larger number of parameters may result in a higher peak likelihood, the extra dimensions greatly increase the volume of the model space, the majority of which will correspond to very low likelihood (Sambridge *et al.*, 2006). Essentially, a more complex model may achieve a better fit to the data, but there are many more ways for it to fail to do so. On the other hand, calculating the evidence can be challenging in the majority of situations where the integral in equation (5) cannot be evaluated analytically, and especially where the number of parameters is large, due to the difficulty in sufficiently sampling the likelihood surface with numerical procedures.

There are various methods for estimating the evidence where an analytical solution does not exist. The most straightforward would be to replace the required integral with a sum, averaging the likelihood over a large number of samples from the prior, but convergence would be impracticably slow in all but the simplest of models. If posterior samples are generated, e.g. through a Markov Chain Monte Carlo model inference procedure, these can be utilised to calculate the evidence by taking their harmonic mean (Kass and Raftery, 1995); however the convergence may be equally poor or even worse with this method (Sambridge *et al.*, 2006). Another approach is to estimate the posterior density, e.g. by using Laplace's method (Wasserman, 2000), and then evaluate posterior, likelihood and prior density at a single set of parameters in order to obtain the evidence via equation (4).

A more common approach to calculating evidence is to use Nested Sampling (Skilling, 2006). This method achieves greater efficiency by reducing the dimensionality of the required integration to a single variable, the prior mass. Using a progressively more constrained set of samples, the enclosed prior mass is shrunk by a particular (known) factor at each iteration, while concurrently the likelihood corresponding to each removed 'layer' is evaluated. The numerical integration of equation (5) is then one-dimensional and straightforward. The iterations continue until a convergence criterion has been reached.

Bayes Factors have been used widely in model selection in different areas of science, from Biology (Aitken and Akman, 2013) to Cosmology (Feroz and Hobson, 2008, Shaw *et al.*, 2007), Acoustics (Jasa and Xiang, 2012, Escolano *et al.*, 2012), and Geophysics (Elsheikh *et al.*, 2013), with Nested Sampling being by far the most popular evidence evaluation method. To our knowledge, they have not yet been used in the selection of statistical models of seismicity.

Runs and Kolmogorov-Smirnov tests

These standard statistical tests are based on a null hypothesis of a Poisson process, which may be rejected at some chosen significance level depending on the calculated p-value.

The Kolmogorov-Smirnov one-sample test evaluates the probability that a specified distribution function is the underlying distribution from which a given sample has been generated, and rejects this null hypothesis if the probability is smaller than a chosen significance threshold (Gibbons, 2011a). In the case of a Poisson process, this can be used to test whether the intervals between events are exponentially distributed.

The' runs test' explicitly uses a null hypothesis of a Poisson process. It is designed to detect autocorrelations in the data by looking at the number of distinct 'runs' of successive values within the data that are above or below the mean value, and evaluating the probability of that number occurring

under the null hypothesis of a Poisson process (Gibbons, 2011b). Either a small or an abnormally large number of runs would indicate the presence of autocorrelations.

These tests have both been used on earthquake data to detect non-stationary behaviour (Mulargia and Tinti, 1985, Lombardi *et al.*, 2010, Wang *et al.*, 2010). For seismicity that is well represented by a stationary ETAS model, the residual point process (RPP) of Ogata (1988) will be a Poisson process of unit rate, making the RPP a suitable event series on which to perform these tests. The residual point process is created by integration of the conditional intensity:

$$\Lambda(t) = \int_{T_1}^{t} \lambda(s|H_s)ds \tag{6}$$

where λ is the ETAS conditional intensity function (equation (1)). Alternatively, the seismicity can be declustered to remove aftershocks and leave only the occurrence times of events that are deemed to be part of the background or 'seeding' process, for example using Zhuang's (2002) stochastic declustering.

Z-score and similar metrics

Another class of changepoint metrics uses the Normal distribution to assess the significance of deviations from a stationary Poisson process after a specific point in time, through a Z-score. Again, for seismicity, the residual point process or a declustered background process may be tested with these.

The expression for the test statistic is derived by constructing a Poisson process as an arbitrary number of consecutive time windows or bins of width δ, in each of which, the distribution of the number of events is Poissonian, with mean $\mu \times \delta$, where μ is the event rate. The number of events in each bin is an independent draw from the distribution. The mean number of events per bin, then, by the central limit theorem, is normally distributed, with a standard deviation $\sigma/\sqrt{n}$, where n is the number of bins, and σ is the sampling standard deviation for the bin count. Because the distribution of

bin counts is Poissonian, this sampling standard deviation is $\sigma = \sqrt{\mu\delta}$. The expression for Z, the test statistic for comparing the mean Poisson rate in the bins with the theoretical rate, is:

$$Z = \frac{(\hat{\mu} - \mu)\delta}{\sqrt{\mu\delta/n}} \quad (7)$$

Eliminating the arbitrary number and width of the bins by substituting $n = \Delta_a/\delta$, where Δ_a is the duration of the potential deviation after the changepoint, and evaluating the measured event rate $\hat{\mu}$ as the number of events N_a in the deviation divided by Δ_a, gives:

$$Z = \frac{N_a - \mu\Delta_a}{\sqrt{N_a}} \quad (8)$$

which will have a standard normal distribution under the null hypothesis.

Matthews and Reasenberg (1987) expressed scepticism about the validity of the Normal approximation and its apparent independence of the length of data. However, the validity depends on the number of bins, which is purely hypothetical since bins are never actually constructed, and this number can be regarded as arbitrarily large. Figure 1 compares, for different values of N_a and Δ_a, (1) the probability of those values derived from calculating Z and converting it to a tail probability, with (2) the Poisson probability $= 1 - \sum_{i=0}^{N_a} \frac{e^{-\mu\Delta_a}(\mu\Delta_a)^i}{i!}$, both under the null hypothesis of a Poisson process with rate $\mu = 1$. These are a reasonable match showing that indeed the standard normal approximation for Z under the null hypothesis is a good one.

Commented [ST1]: page 34-35

Alternative metrics in the same vein include Matthews and Reasenberg's (1988) β, a variant of Z above which deals with the slightly more elaborate situation of an anomalous period bounded in time, taking the form of a step change in μ followed by a step change back to its original value. Ogata (1992) uses a statistic ξ, which introduces a correction to the denominator in Z above to account for uncertainty in the 'base' (or null hypothesis) Poisson rate due to it being estimated from a finite sample. Effectively this enlarges the standard deviation, particularly at small sample sizes. Habermann (1983) presents the equivalent Z score for a difference between two means, i.e. the mean

rates before and after a changepoint. This is based on a null hypothesis of stationarity (equality of observed rates before and after the changepoint), unlike the simple Z above which compares an observed mean rate with a theoretical (null hypothesis) value. Habermann's Z is given by:

$$Z = \frac{N_a\Delta_b - N_b\Delta_a}{\sqrt{N_b\Delta_a^2 + N_a\Delta_b^2}} \quad (9)$$

where the subscripts 'a' and 'b' refer to after and before the changepoint, respectively. This method is also widely used in detecting non-stationarity in seismicity data (Wyss and Wiemer, 2000, Katsumata, 2011, Marsan and Nalbant, 2005).

In terms of detecting a changepoint, these Z-type metrics, unlike the information criteria, evidence, Kolmogorov-Smirnov or runs methods, require the potentially anomalous period – and hence the changepoint time – to be specified in advance of the evaluation. The identification of a changepoint for use in a Z-type test entails extra considerations which we discuss in a later section.

Note that deriving a theoretical long-term rate from data, as required for the simple Z test, involves a choice of whether to include the potentially anomalous period in the rate calculation. Clearly, including it will give a higher theoretical rate and make the changepoint harder to accept. Similarly, for analysis of correlated seismicity, for example using the ETAS model, there is a choice about whether to base the RPP on parameters inverted from the whole catalogue, or from only the period prior to the changepoint. If the null hypothesis is one of no change, then it seems logical that the estimation of the rate and the RPP should be based on the whole catalogue.

3. Testing changepoint detection methods on simulated samples

We test the changepoint detection methods described in the previous section using simulations of the Poisson and ETAS models as follows. We simulate 500 realisations of each case examined, upon which to perform the analysis. Each realisation has 1000 events, in which the first 500 events are generated using a background rate μ_1 and the second 500 events are generated with a rate $\mu_2 = \upsilon \times$

μ_1, where v ranges from 1 to 2. For the ETAS simulations, we also include an extra 1000 events prior to the start of the sample (using the rate μ_1) as 'history', to be used in evaluating the likelihood for the sample of interest; this is to reduce the possibility of wrongly ascribing aftershocks to the background process when the parent event occurred prior to the start of the sample. For the Poisson tests, we give μ_1 the value 0.1 in every case (the results – as one would expect – are insensitive to the choice). For ETAS, the timescale of the aftershock rate decay does not change with the rate of the independent events, and the relative timescales of these two processes have been shown to affect the statistics of the time series and the detectability of aftershock parameters (Touati *et al.*, 2009, Touati *et al.*, 2011); the effect of the parameter μ_1 thus has to be explored. We choose three different values for μ_1: 0.01, 0.1 and 1. A Gutenberg-Richter frequency-magnitude model is used, with parameter b=1; this and the other parameter values used (A=10, α= 1, c= 0.01, p = 1.2) are constant throughout the simulation and result in a branching ratio of 0.88, which is in the typical range for tectonic seismicity.

The question to which we seek an answer is: given a set of 1000 events, how successful are the methods described in the previous section at (1) rejecting the null hypothesis of a stationary process when there is a change in μ within the data, and (2) declining to reject the null hypothesis when the data is stationary? We essentially measure the rates of type 1 and type 2 statistical errors (Dimer de Oliveira, 2012). In some ways this is a limited question; real seismicity studies often deal with multiple changepoints, longer catalogues, and potential changepoints that are far from the centre of the data set. However, the insights gleaned on the efficacy of these methods for this particular situation can guide a strategy for other situations; they also provide an example methodology for evaluating the efficacy, that may be replicated for other situations. Although we test the methods on simulations with an increased rate following a changepoint, the results are statistically equivalent to the corresponding case of switching to a quiescent period (none of the tests are constrained only to detect an increase in rate), and are thus applicable to such situations also.

Carrying out the tests with AIC and BIC requires fitting the appropriate model (Poisson or ETAS) to the simulation. We use maximum likelihood to fit both (1) the stationary model, and (2) a

'changepoint' model, in which μ's value changes from μ_1 to μ_2 at a changepoint t_c (in time), and all other parameters (in the case of ETAS) are constant throughout the sample. The changepoint model therefore has seven parameters: μ_1, μ_2, A, α, c, p and t_c. The maximised log likelihoods can then be compared through AIC and BIC to discover which model is preferred. A significance threshold of 4 for the AIC difference is applied. BIC does not have a clear rule of thumb for significance, but we use a threshold of 2, by analogy with the log Bayes Factor, which is deemed 'decisive' for values greater than 2 (Kass and Raftery, 1995).

Initial attempts to calculate the Bayesian evidence via a simple numerical evaluation of equation (5), by sampling from the priors and averaging the likelihood, were not successful. After 2,000,000 samples, the evidence had still not converged, due to the size of the model space (even with the relatively small number of parameters in the ETAS models) relative to the size of the peak region and its much higher likelihood. Nested Sampling was found to be a good alternative, using the code of Sivia and Skilling (Sivia, 2007) as a basis.

The Nested Sampling procedure requires an initial number of parameter samples η to be generated from the prior; we use $\eta = 100$ samples here, from a uniform prior. At every iteration, the sample with the worst likelihood is discarded and replaced by a new sample with a better likelihood, after including its contribution to the evidence. In this way the evidence is computed as the sum of contributions from approximately concentric bands of the model space. The most challenging and time-consuming aspect of the procedure is the selection of a new sample from the prior under the constraint that its likelihood must be higher than a particular value. As the algorithm proceeds, typically a large number samples must be tried and rejected at each iteration. Ellipsoidal sampling (Shaw *et al.*, 2007) exploits the (usually) ellipsoidal shape of the likelihood contours, by finding the spanning ellipsoid for the existing set of samples within the parameter space, enlarging it slightly, and obtaining a uniform sample from within this ellipsoid (assuming the prior is uniform) as a candidate new sample. We implement this method, with an 'enlargement factor' for the ellipsoid of 1.1. For a changepoint model with stationary data, however, the likelihood surface is highly non-ellipsoidal due

to the lack of constraint for the changepoint parameter in the data. The changepoint has a higher likelihood of being very near either the beginning or end of the data set, presumably due to the ease of fitting the very small number of events captured at one side of the changepoint with almost any event rate. This results in two 'arm'-like structures in the likelihood surface, and hence, highly inefficient sampling using the ellipsoid approximation. We tested two ways of handling this situation: (1) truncating the prior on the changepoint so that it was restricted to be within the middle 600 events of the sample, thus eliminating the 'arms'; (2) employing a cuboidal sampling regime with three cuboids, one for each of the 'arms' and one central one, that neither overlap nor leave gaps between their boundaries. Both of these interventions were found to greatly improve the efficiency, and the results were not significantly different between the two methods. The results we present here are using the restricted prior on the changepoint.

Nested Sampling makes convergence of the evidence feasible, but the computation time is still lengthy, particularly for ETAS, where a single nested sampling run for the changepoint model takes over a week on a desktop computer. Testing the procedure on 500 realisations of ETAS was not possible due to this computational cost, but results for a single realisation of each value of μ_1 and two different values of v (1 and 2) were obtained as an indication. For Poisson simulations, it was possible to process the full 500 realisations of each case. The preference for a changepoint model can be deemed significant for a log Bayes Factor of at least 0.5, and 'decisive' for values greater than 2 (Kass and Raftery, 1995); we use a threshold of 2 for accepting a changepoint.

For the tests that use a Poisson process as a null hypothesis, we apply these to ETAS simulations by producing both the residual point process (Ogata, 1988) and a stochastic declustered version of the sample (Zhuang *et al.*, 2002), both of which should take the form of a stationary Poisson process under the null hypothesis. In order to incorporate the parameter inversion error that would be present in a real situation, we optimise the parameters as if they are unknown. This is done over (1) the first 500 events, i.e. the period prior to the potential changepoint, and (2) the whole 1000-event sample, to analyse the consequences of this choice.

In our implementation of the simple Z test, which compares the average rate in a particular time period with a theoretical long-term value, the theoretical rate is estimated both possible ways: from the first 500 events, and from the whole 1000 events. In all our Z tests, we take the pre-specified location of the changepoint to be the 501st event, deferring discussion of the issues around making such a choice to the next section. We use the 95% confidence bounds of the standard normal distribution, or a deviation of ±2, as the threshold for significance. For the Kolmogorov-Smirnov and runs tests we also reject the null hypothesis if the value of the relevant statistic falls outside of the 95% confidence bounds.

4. Results of simulation tests

Poisson simulations

The results of testing the above changepoint detection methods are collated in **Figure 2**(a) which shows the fraction of realisations for which the changepoint was accepted as a function of v. Figure 2(b) is a cartoon to help identify key features in Figure 2(a); in particular, the rate of false positives when there is no rate change, and the threshold at which 95% of true rate changes are accepted. Figure 2(c) summarises the performance of each method with regard to these metrics. The simple Z (with the long-term rate derived from the first half of the dataset) and the runs test have the greatest rates of false positives; they tend to fail on different datasets (see Appendix).

AIC, BIC and the Bayes Factor all produce relatively high changepoint-acceptance rates in cases where a changepoint does exist. BIC and the Bayes Factor both give very low acceptance probabilities for a stationary process. AIC rejects the stationary Poisson process slightly more often, giving a lower rate of type 2 errors but a slightly higher rate of type 1 errors. BIC slightly outperforms the Bayes Factor throughout; this is somewhat surprising given the more rigorous approach in calculating the full Bayesian evidence. There is some subjectivity in the precise choice of acceptance threshold for the Bayesian methods; in practice we would recommend calibrating the thresholds to simluations

which are conditioned on the problem of interest to help inform the choice of threshold. However, the general results are robust.

The remaining tests do not rely on parameter inversions on the Poisson simulations, and so merely reflect inherent properties of the Poisson process in the situations presented.

The runs and Kolmogorov-Smirnov tests, which are designed to reject a stationary Poisson process on grounds of temporal clustering or the poor fit of an exponential distribution for the time intervals between events, respectively, are relatively poor at rejecting the null hypothesis, in this context of a Poisson process with a change in rate half-way through the data. Even with a doubling of the rate, the rejection probabilities are only around 0.55 and 0.3, respectively. These two tests are the most generic of the methods evaluated here, and do not incorporate the concept of a changepoint, so perhaps it is not surprising that they are outperformed by other, more specific tests.

For the Z tests, the Habermann Z gives a type 1 error rate of about 0.05, which is as expected given the 95% confidence bounds. These rejections represent null-hypothesis realisations that have a spurious increase or decrease in the average rate after the changepoint – an effect that appears significant by chance. The rejection rates when using data with a changepoint (i.e. type 2 errors) are lower than those with AIC, BIC or the Bayes Factor.

The simple Z test naturally suffers from the inaccuracy in the theoretical long-term rate, whether this is derived from the first 500 events or the whole 1000 events: the significance of a deviation from this rate after the changepoint is over- and under-estimated, respectively, in comparison to Habermann's Z. We would expect the results for simple Z to converge towards Habermann's Z given a longer event history prior to the changepoint; also, we note that Ogata (1992) provided a modification to the simple Z to account for the uncertainty in the long-term rate when the event history is short.

ETAS simulations

In moving from a Poisson process to one that also includes triggered earthquakes (the ETAS model), we find that the introduction of aftershocks generally leads to substantially fewer changepoint detections, i.e. more type 2 errors. Aftershocks themselves are a form of clustering or non-stationarity, which presents the challenge of accounting both for the aftershock process (under sometimes considerable uncertainty) and any additional non-stationarity caused by a change in the seeding rate, μ, as two separate effects. We have previously shown that the degree of temporal overlap of separate aftershock sequences is an important determinant of the accuracy with which the aftershock statistics can be inferred using maximum likelihood (Touati *et al.*, 2011); overlap results in a masking effect whereby the signature of the aftershocks in the time intervals between events is lost. Thus it is difficult to distinguish between aftershocks and independent events when the independent event rate (μ) is high. Through the current simulation exercise, we provide a quantitative analysis of the difficulty of distinguishing between aftershock clustering and a change in the basic background or random seeding rate.

In agreement with the Poisson case, the performances of AIC and BIC on ETAS were found to be comparable, with AIC giving slightly more frequent changepoint acceptances. We neglected to assess the runs and Kolmogorov-Smirnov tests for the ETAS simulations since the results had been so poor for the Poisson case. For the Z tests, we again found that the simple Z test with the two different estimations of the theoretical rate (the average rate before the changepoint, and the average rate in the whole sample) gave two different results, and that the Habermann Z results were in the middle. We found no large differences in the results when using optimised parameters from the whole of the sample versus from only the period before the changepoint, in creating the RPP or declustered sample. We also found that using the RPP or a declustered sample gave comparable results.

Figure 3 compares the performance of Habermann's Z (tested on the residual point process obtained from an inversion of ETAS parameters for the whole of each sample) with that of AIC, and also shows the effect of the absolute value of the background rate, μ_1, in both of these tests. In agreement with the Poisson case, Habermann's Z test performs broadly similarly to AIC, but produces more type

Commented [ST2]: page 36

1 errors (more than the expected rejection fraction of 0.05 when $\nu = 1$). Interestingly, changepoint
2 acceptance probability seems to be a decreasing function of μ_1 for AIC. This effect can be explained
3 in terms of our earlier results on the masking of aftershock statistics: higher μ would tend to make the
4 inversion of parameters, for both the stationary and the changepoint models, less accurate. The
5 likelihood functions of both would be flatter, the difference in likelihood smaller, and thus the
6 difference in AIC less likely to be positive. The Z test, by contrast, is dependent only on fitting
7 stationary ETAS. Since the result of this becomes biased towards a lower branching ratio when μ is
8 high (Touati *et al.*, 2011), the conversion to a residual point process should result in a more dramatic
9 change in slope at the changepoint, as the aftershocks in the data cannot all be accommodated by the
10 model as aftershocks. However, inter-event times are more exponentially distributed at high μ anyway
11 (Touati *et al.*, 2009), which perhaps mitigates this effect in the Z test.
12
13 For the Bayes Factor, the computation time involved in performing the nested sampling precluded a
14 Monte Carlo analysis on 500 realisations, but the log-evidence for a single realisation of each case is
15 given in table 1, along with its standard deviation. The standard deviation is due to uncertainty in the
16 approximation used for prior mass at each iteration of the sampling; this source of error dwarfs the
17 numerical integration error, and is given by $\sqrt{H/\eta}$, where H is the 'information', a measure of how
18 informative the data are relative to the prior. These results, interestingly, are not consistent with the
19 general pattern in Figure 3 of type 2 errors being more common than type 1. The preference for a
20 changepoint model can be deemed significant for a log Bayes Factor of at least 0.5, and 'decisive' for
21 values greater than 2 (Kass and Raftery, 1995); all of the acceptances in table 1 are therefore decisive.
22 The values of the standard deviation for log evidence, however, admit the possibility that some of
23 these are *not* decisive; caution must also be used when inferring a general pattern from a single
24 realisation. However, it does seem that the method may accept more spurious changepoints than
25 Habermann's Z and AIC, and more so for larger μ_1.
26

We have a couple of comments to make on the results of these simulation tests and their implications. Firstly, the assumption of a standard normal distribution for Z may need to be relaxed in the case of ETAS, where even in the null hypothesis of a stationary ETAS model, error in the parameter inversion required for transforming the data to a residual point process can propagate to a null Z distribution that is wider than expected. Figure 4 shows the histograms of Habermann's Z for stationary ETAS simulations, for the three different μ_1 values, along with the expected histogram based on the standard normal. For high μ_1, the distribution is significantly different from the expectation, with fatter tails. A t-test might therefore be more appropriate in situations with a high background rate. This would reduce the number of type 1 errors associated with this test, which Figure 3 shows is indeed slightly higher for high μ_1.

Commented [ST3]: page 37

Secondly, the null hypothesis is that Z has a standard normal distribution for any *given* sample; that is, any period of data selected at random – which is precisely the situation represented by these simulation tests. A changepoint in a catalogue, however, is not generally identified by selecting a single candidate changepoint at random and rejecting or accepting it. Often the procedure is to scan through the catalogue and look for the most extreme value of Z, which is asking a different question than that asked by the simulation exercise above: is there a significant changepoint anywhere in the whole catalogue? The answer again depends on what can be expected under the null hypothesis, which in this case must also depend on the length of the catalogue, because a larger number of extreme events are to be expected in a longer catalogue. The null distribution for Z is therefore not the standard normal whenever scanning or optimising for a changepoint is part of the procedure. This is pointed out by Matthews and Reasenberg in a comment on the Habermann method (Matthews and Reasenberg, 1987).

Naus (1982) provides an accurate approximation to these so-called scan statistics, for the Poisson probability of a given number of events occurring in a given time window somewhere within a larger period. This would find application in a slightly different situation from the simple changepoint identification: for example, the identification by eye of a surprising cluster (or quiescent period)

within a record of events, and the desire to know whether it is extreme enough not to have arisen by chance within a record of that length, as part of a long-term Poisson process. If the surprising period occurs at the beginning or end of the record, however, then it is equivalent to the single-changepoint problem.

Recognising the effect of scanning through a catalogue for extreme values of the test statistic, Ogata (1992) and Matthews and Reasenberg (1988) use an empirical null distribution for their changepoint-detection test statistics, derived from stationary Poisson simulations. Habermann (1983), on the other hand, uses the standard normal. We can compare the performance of Habermann's Z with the Poisson scan statistics (using Naus's approximation) in assessing the significance of an apparent cluster in a long catalogue, as follows. We construct an apparently anomalous period of 50 days in which 70 events occur, and an event history preceding that in which we assume the average event rate is found to be 1 event per day. Whatever the length of the history, we set the candidate changepoint to be at the start of the 50-day cluster, making the assumption that an optimisation would identify this as the most extreme changepoint. For the Poisson statistics, we assume that the theoretical rate is considered to be 1 event per day, i.e. the rate prior to the changepoint, noting that if the whole-catalogue average rate was used instead, this would result in a higher probability (and more so at shorter catalogue lengths). Figure 5 shows the analytical probability of this cluster as a function of catalogue length, using both Poisson scan statistics and Habermann's Z. Z has been converted to the probability in the upper tail of the standard normal, i.e., the probability of this increase in rate or a bigger one; correspondingly, the Poisson probability is of 70 events or more within 50 days. The 95% confidence limits for the null hypothesis imply that a probability of less than 0.025 would make the cluster significant in a two-sided test. With the shortest event history of 50 days, the subsequent cluster is not found to be outside the 95% confidence limits by either Habermann's Z or the Poisson probability. For Z, the probability then drops into the significant range when an increasing catalogue length (or event history) is considered – due to a smaller uncertainty in the event rate prior to the changepoint – and quickly becomes indifferent to catalogue length. However, for the Poisson scan statistics, the probability of the cluster grows considerably with catalogue length and approaches 1, such that the occurrence of

Commented [ST4]: page 38

this hypothetical cluster is not considered surprising at all in a long enough catalogue. Because of their equivalence with the simple Z, the Poisson scan statistics results can be assumed to be equivalent to the probability obtained from a simple Z that is optimised and evaluated against an empirical null distribution for the situation, as done by Ogata (1992) and Matthews and Reasenberg (1988). A similar result should be expected using Habermann's Z combined with an appropriate empirical null distribution. Even though Habermann's Z explicitly depends on the length of the event history, it does not do so in a way that accounts for the scanning, when combined with a standard normal assumption for the null distribution.

5. Global large earthquakes data

Having tested the efficacy of several changepoint detection methods on simulations, we now turn our attention to a real setting in which these methods could be applied. The assumption that the largest events worldwide are a stationary Poisson process has been called into question in recent years (Bufe and Perkins, 2005), primarily due to the appearance of a temporal clustering of megaquakes from 2004 to the present. Given this apparent changepoint, global megaquakes are a good case study upon which to apply the best-performing techniques for changepoint detection as ascertained in this paper, and to illustrate the usefulness of our findings.

Global catalogues have a relatively high completeness threshold compared with those of regional catalogues. When earthquake catalogues are truncated at higher and higher minimum threshold magnitudes, the effect is to make the events appear more independent – partly because, since the majority of aftershocks are smaller than their parent event, the events remaining after this truncation are indeed more likely to be independent; and partly because in some cases, two events that are part of a common triggering sequence may be deemed independent after the removal of intermediate related events below the threshold, leading to a lowering of the apparent branching ratio (Saichev and Sornette, 2006). Selecting only the very largest worldwide earthquakes makes the Poisson assumption reasonable, and the more so the higher the inclusion threshold is raised, but it remains an approximation. Some declustering is often applied when analysing global data, using criteria of

temporal and spatial proximity to infer causally related events; for example, using the values given by Gardner and Knopoff (1974).

Some studies on global megaquakes have looked at the numbers of events in equal-length time windows. Examples include (Daub *et al.*, 2012), who ran Poisson simulations to compare the occurrence frequencies of such numbers in the data with those in the null hypothesis. Bufe and Perkins (2005) and Shearer and Stark (2012) both look for occurrence frequencies of whole clusters (or quiescences) in simulations, but construct these 'anomalies' under different parameters and obtain different results. Shearer and Stark (2012) used the Poisson dispersion test and the multinomial chi-squared test as discriminants, and point out that every realisation of a random process contains some feature that is highly unlikely, and the more specifically the feature is described (e.g. through optimisation of some parameter for the most extreme result), the more unlikely it becomes. The Kolmogorov-Smirnov test has also been used to see whether a Poisson process should be rejected (Michael, 2011).

Here we apply the more successful of the changepoint detection methods tested in sections 2-4 to global large events, using data from the recent International Seismological Centre-Global Earthquake Model (ISC-GEM) worldwide catalogue (Storchak *et al.*, 2013). This is a unique catalogue extending from 1900 to 2009, giving homogeneous moment magnitudes and uncertainties, based on hundreds of data sources and uniform techniques. We supplement this with post-2009 moment magnitude data from the Centroid Moment Tensor (CMT) catalogue (Ekstrom *et al.*, 2012), up to the 15th of October 2014. The period 1st July – 15th October 2014 comes from the 'Quick CMT' catalogue of the most recent events.

The completeness threshold for the ISC-GEM catalogue is 7.0 for the period 1918-2009; the earlier period of 1900-1917 is likely to be incomplete even above that level (Michael, 2014). We therefore analyse data from 1918 onwards, of magnitudes 7.0 and above, at all locations and depths. Thus, including the CMT data, we have 96.79 years of data. We look at data above four different magnitude

thresholds: 7.0, 7.5, 8.0 and 8.5. In each case we analyse the entire data, but also compare this with a declusered subset of the data, obtained by applying the Michael (2011) criteria to identify triggered sequences of events and remove all but the largest event from each identified sequence. Declustering using ETAS would be inappropriate for the whole globe due to wide variation in tectonic styles and associated parameter values. However, the insight of Saichev and Sornette (2006) that triggering relationships tend to be broken when increasing the magnitude threshold is certainly relevant for the whole Earth. Taking this into account, we perform the Michael (2011) declustering on data of magnitude 5.0 and above before applying the desired (higher) thresholds to the result, in order to minimise the number of related events that are mistakenly taken to be independent. This is found to result in the removal of a significantly larger number of events compared with applying the threshold first: for example, 179 events (or 16%) at magnitude 7.0 and above are removed, compared with only 94 events (8%) in the same range when applying the declustering algorithm to data that is first truncated below magnitude 7.0.

The numbers of included events for each analysis case are given in Table 2. Figure 6 shows the average event rate as a function of time through the catalogue, for each threshold, with and without the declustering. The average is calculated in windows of 5 years from 1918 to 2013, and then with a final window of 1.79 years. (The windowing is a reasonable compromise between smoothness and detail, but is purely used for visualisation.) There is a period of higher activity around 1970, an apparent relative quiescence from 1980 to 1990, and then an increasing trend from 1990 onwards. These trends are still present to some extent in the declustered data. Thus, it seems reasonable to ask whether a change in event rate has occurred somewhere in the recent past.

Commented [ST5]: page 39

We fit both the stationary and the changepoint Poisson models to each set of data using the maximum likelihood method. The changepoint model was more difficult to fit as it appeared to have a complex likelihood surface without a clear peak; a variety of different starting values for the changepoint time were used in order to be confident that a global maximum was found. The results of the parameter inversion are given in the third and fourth columns of table 2. The maximised likelihoods are then

used to compute the difference in AIC and BIC between the two models, with positive values indicating preference for a changepoint model.

We also apply the nested sampling algorithm to compute the Bayes Factor; again, positive values indicating preference for a changepoint. The Habermann Z is computed using the changepoint time inverted in the Poisson changepoint model. This changepoint time is also used in calculating the probability of the data occurring after the changepoint using the Poisson scan statistics of Naus (1982); the theoretical long-term rate used in the latter calculation is the average event rate during the whole data set, from 1918.0-2014.79, reflecting a null hypothesis of no rate change. For some of the cases with lower a magnitude thresholds, this could not be calculated due to the factorial of the number of events in the 'cluster' – required as part of the calculation – being too large for computational representation.

The results of all tests are presented in table 2. The post-changepoint rate μ_2 inferred by fitting the changepoint Poisson model is greater than μ_1 in every case, confirming the observation of an apparent increase in recent years. However, the changepoint identified in the model varies by more than ten years between the different data sets. In terms of its acceptance, AIC tends to accept a changepoint at low threshold and reject it at higher thresholds; the declustering generally decreases the favourability of the changepoint, which is never accepted decisively for declustered data. BIC does not show the same variation with the threshold value, perhaps because of its accounting for the number of data points; it always rejects a changepoint more strongly for declustered data, and it rejects a changepoint for all but the $M \geq 7$ non-declustered series. The Bayes Factor rejects the changepoint in all cases except for non-declustered data with threshold 7.5 (but this latter case is not decisive). In Figure 2(a,c), it can be seen that the Bayes Factor tends to reject true changepoints more often than BIC does, but that these methods have a comparable, low probability of accepting a changepoint in stationary data. The similarly low changepoint acceptance rates of BIC and the Bayes Factor for this global data, then, might be taken as an indication that the data is stationary.

For Habermann's Z, which is intended to have a standard normal distribution under the null hypothesis, the significance level for 95% confidence is ±2. Thus Habermann's Z accepts the changepoint in all cases except for the data with threshold 8.5 and the declustered data with threshold 8.0. It must be borne in mind, of course, that evaluating the significance in this way is misleading as it does not take into account the increasing occurrence frequency of extreme events with data length. The values given here can only be taken as an upper limit. When using the scan statistics with the Poisson probability to include this dependence on data length, the probabilities of the 'cluster' periods following the changepoints are in all cases substantially higher than the significance probability level of 0.025.

As stated earlier in this section, we found that significantly more events were removed when declustering the data *prior* to applying a magnitude threshold than when declustering *after* the truncation. This is because truncation below a magnitude threshold 'hides' some triggering relationships that may exist between events that remain in the catalogue, relationships that no longer appear to exist after removal of the intermediate (smaller) events in the sequence. Indeed, when the changepoint detection methods are applied to data that is declustered after applying a threshold of 7.0, the results (not shown) are more similar to those for the non-declustered data. This poses the question: have all aftershocks been removed by declustering a catalogue of magnitude 5.0 and above; or would an even lower threshold be needed to successfully identify all triggering relationships prior to an analysis that seeks to exclude these from any conclusions about anomalous temporal clustering? According to Helmstetter (2003), it depends on the way aftershock productivity and occurrence frequency both scale with magnitude, and it may well be the case that triggering is driven by the smallest earthquakes; in this case, a threshold much lower than 5.0 (if global data were plentiful enough) would probably remove significantly more events and change the conclusions.

In summary, when comparing the results of several different changepoint methods on global large earthquakes, and having tested their efficacy in previous sections, we see no strong evidence that the basic rate of large events worldwide has increased in recent years. Of the cases we examined, none

had a changepoint accepted by all methods. Given the sensitivity of the methods to changes in event rate for Poisson processes, demonstrated in Figure 2, we can conclude that if there is a change in event rate, it is relatively small. We expect that the data declustered at magnitude 5.0 and above will still contain some triggered events, and thus a small fluctuation in event rate (if it did exist) could easily be attributed to this.

6. Discussion and conclusions

In our results for the tests on Poisson simulations, the type 1 error rate appears to be higher for AIC and Habermann's Z than for BIC and the Bayes Factor; we conclude that strong rejection by AIC or Habermann's Z could therefore be quite decisive.

We did not obtain good results using the Kolmogorov-Smirnov or runs tests, with Poisson simulations. There were a large number of type 2 errors, and for the runs test, type 1 also. This is likely to be because these are more general tests not dealing with changepoints specifically but with general deviations from a Poisson process.

Habermann's Z or Ogata's ξ are preferable to the simple Z, because of the inaccuracy in any estimate of the long-term rate for the latter. However, it is very important that the length of the catalogue is taken into account when assessing the significance of a Z value. An extreme event (such as a 'cluster') in a very long data set is actually an expected outcome in a random process. Choosing a changepoint by eye should, in our opinion, be viewed as the result of an informal scanning or optimising procedure. An empirical null distribution for Z should be computed for the specific situation tested, rather than assuming a standard normal distribution.

That having been said, it is still the case that every realisation of a random process contains some feature or other that is statistically highly unlikely. A highly unlikely sequence or pattern is not always an imperative to reject the null hypothesis. The more specifically an 'anomalous' feature of a data set

is described, the more unlikely it appears to be (Shearer and Stark, 2012); very narrow descriptions of highly anomalous occurrences, that are a result of having seen the data, should be avoided.

Scan statistics are often easier to compute than a null Z distribution, and should give equivalent results to using the simple Z with an empirical null distribution – although for long clusters, the factorial of the number of events may be large enough to make this method impractical. Poisson scan statistics are a good alternative to Habermann's Z or Ogata's ξ when there is a long history and a short cluster.

When using the ETAS model to convert a seismicity catalogue into a Poisson process prior to using a Z-type metric, i.e. computing the residual point process or declustering with ETAS, the empirical null distribution for Z should ideally be derived not from Poisson simulations but from ETAS simulations upon which the inversion of model parameters and calculation of the RPP or background process has been performed, in order to account for error in that process. This would obviously be more time-consuming. In general, we found changepoint detection in ETAS-type data to be much more challenging, with a higher rate of type 2 errors in particular (Figure 3).

Using AIC, BIC and the Bayes Factor to choose between competing models would be a good complementary strategy to the above methods which require a changepoint to be specified. The changepoint can be one parameter in a model, which incurs a penalty (either explicitly in AIC/BIC, or implicitly in the Bayesian evidence) analogous to the 'penalty' imposed by use of the empirical null distribution for a Z-type metric which would be wider than the standard normal. We found that the results using the Bayes Factor were surprisingly poorer than those for BIC, for the Poisson simulation tests we carried out. Given the computational cost of evaluating the Bayes Factor, particularly for more complex models than Poisson (such as ETAS), we would suggest that AIC and BIC are sufficient, provided the global maximum in the likelihood function can be found with confidence.

When declustering data to remove normal aftershock triggering processes before testing for deviations from a Poisson process, it is important to use all available data, including events smaller than the

completeness threshold and/or the threshold of interest. Our analysis of global data has shown that the results can be significantly altered by declustering data that is already magnitude-truncated, due to the incomplete removal of aftershocks in that case. This is further evidence that smaller events play an important overall role in triggering (Helmstetter, 2003).

Changes in magnitude determination methods can give rise to spurious rate changes on the order of those simulated in this paper; magnitude stability would need to be established before reaching any conclusion on a change in the basic rate of earthquakes.

Our analysis provides no strong evidence that the basic rate of large events worldwide has increased in recent years. Of the cases we examined, none had a changepoint accepted by all methods. Given that the data was declustered at magnitude 5.0 and above, a small fluctuation in event rate (if it did exist) is likely to be a result of incomplete removal of aftershocks.

Our results are appropriate for models where the background rate has changed from one stationary state to another. We have not yet considered multiple change points or more transient acceleration or decelerations in forcing. In fact there has been a suggestion of two proposed clusters in global mega-earthquakes, one starting in 1960 and one starting in 2004. Accordingly, Michael (2011, test 3 in that paper) included a two-window test on moment release and Benioff strain (test 3 in that paper). Nevertheless our results may prove useful in the analysis of change points in places such as Oklahoma (e.g. Llenos and Michael, BSSA, 2013) where the main extra forcing term appears to be a sudden change in the total rate of re-injection of waste water produced from hydraulic fracturing for shale gas extraction.

R codes for the tests examined in this paper will be published on http://www.corssa.org/.

7. Appendix

Here we briefly examine the question of whether there is any cross-correlation between the different tests in terms of the likelihood of changepoint acceptance. We do this by examining changepoint acceptance by individual synthetic dataset, for the Poisson simulation tests presented in Figure 2.

Figure A.1 shows the occurrence of false positives. The graphic only shows the 109 out of 500 datasets for which false positives were identified in at least 1 test. The tests producing the largest numbers of false positives were the simple Z (with long-term rate derived from the first half of the data) with 78 false positives, and the runs test, with 27 false positives. Broadly speaking, there is a nesting of the failure of the tests: Habermann Z never fails without the simple Z (with long-term rate derived from the first half of the data) also failing; and AIC rarely fails without the Habermann Z and simple Z also failing. However, the runs test mostly gives false positives on different datasets.

We see this pattern broadly echoed in Figure A.2, which shows the occurrence of true positives based on synthetic datasets where μ increased by a factor 1.1 at the changepoint. Again, only the datasets with a changepoint acceptance by one or more tests are shown; this amounts to 295 datasets. The simple Z (with long-term rate derived from the first half of the data) still produces the largest number of changepoint acceptances – this time correctly – at 269. The runs test acceptance rate is almost the same as in the stationary data, at 24. Since there is no increase in the number of positive results for the runs test, we conclude that it is not sensitive to small rate changes.

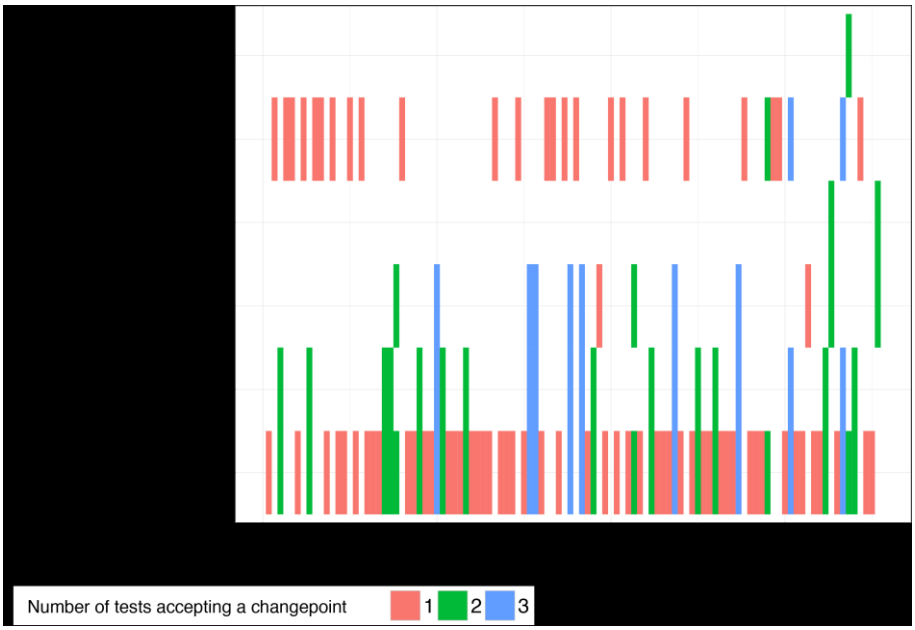


Figure A.1: Occurrence of false positives by synthetic dataset (x-axis) and by test (y-axis) in Poisson simulations when there is no rate change. "Simple Z (first half)" denotes a simple Z test in which the long-term rate was derived from the first half of the data.

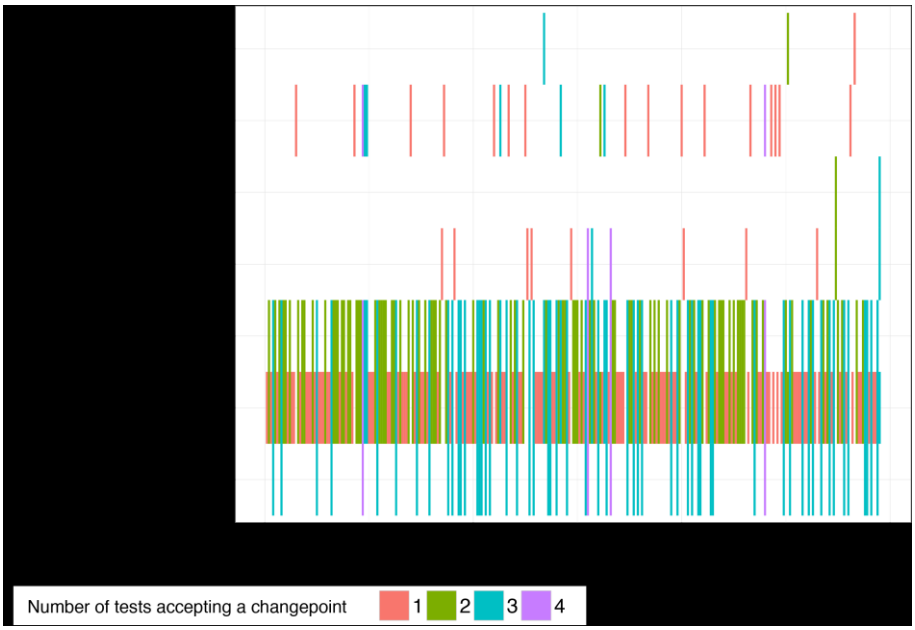


Figure A.2: Occurrence of true positives by synthetic dataset (x-axis) and by test (y-axis) in Poisson simulations when there is a rate change with rate multiplication factor 1.1. "Simple Z (first half)" denotes a simple Z test in which the long-term rate was derived from the first half of the data; "Simple Z (whole)" has its long-term rate derived from the whole dataset.

Acknowledgements

This research was carried out in the framework of the REAKT Project (Strategies and tools for Real-Time EArthquake RisK ReducTion) founded by the European Community via the Seventh Framework Program for Research (FP7), contract no. 282862. ST was funded by REAKT.

MN was funded by a Scottish Government and Royal Society of Edinburgh Research Fellowship.

We are grateful for discussions and advice on Bayesian Evidence from Malcolm Sambridge, and on Nested Sampling from Daniel Tait.

References

Aitken, S. & Akman, O.E., 2013. Nested sampling for parameter inference in systems biology: application to an exemplar circadian model, *Bmc Syst Biol,* 7.

Akaike, H., 1974. New Look at Statistical-Model Identification, *Ieee T Automat Contr,* Ac19**,** 716-723.

Ammon, C. J., Lay, T. & Simpson, D. W., 2010. Great earthquakes and global seismic networks, *Seismol Res Lett,* 81 (6), 965-971.

Bell, A.F., Naylor, M. & Main, I.G., 2013. The limits of predictability of volcanic eruptions from accelerating rates of earthquakes, *Geophys J Int,* 194**,** 1541-1553.

Bufe, C.G. & Perkins, D.M., 2005. Evidence for a global seismic-moment release sequence, *B Seismol Soc Am,* 95**,** 833-843.

Burnham, K.P. & Anderson, D.R., 2002. Model selection and multimodel inference: A practical information-theoretic approach, Springer.

Cowie, P.A., Vanneste, C., & Sornette, D., 1993. Statistical physics models for the spatiotemporal evolution of faults, *J Geophys Res*, 98, 21809-21821.

Daub, E.G., Ben-Naim, E., Guyer, R.A. & Johnson, P.A., 2012. Are megaquakes clustered?, *Geophys Res Lett,* 39.

DeMets, C., 1995. Plate motions and crustal deformation, *Rev Geophys*, 33, 365-369.

Dimer de Oliveira, F., 2012. Can we trust earthquake cluster detection tests?, Geophys. Res. Lett., 39, L17305.

Ekstrom, G., Nettles, M. & Dziewonski, A.M., 2012. The global CMT project 2004-2010: Centroid-moment tensors for 13,017 earthquakes, *Phys Earth Planet In,* 200**,** 1-9.

Ellsworth, W.L., 2013. Injection-Induced Earthquakes, *Science,* 341**,** 142-+.

Elsheikh, A.H., Wheeler, M.F. & Hoteit, I., 2013. Nested sampling algorithm for subsurface flow model selection, uncertainty quantification, and nonlinear calibration, *Water Resour Res,* 49**,** 8383-8399.

Escolano, J., Perez-Lorenzo, J.M., Xiang, N., Cobos, M. & Lopez, J.J., 2012. A Bayesian inference model for speech localization (L), *J Acoust Soc Am,* 132**,** 1257-1260.

Feroz, F. & Hobson, M.P., 2008. Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses, *Mon Not R Astron Soc,* 384**,** 449-463.

Gardner, J.K. & Knopoff, L., 1974. Is the Sequence of Earthquakes in Southern-California, with Aftershocks Removed, Poissonian?, *B Seismol Soc Am,* 64**,** 1363-1367.

Gibbons, J.D.a.C., S., 2011a. 4 Tests of Goodness of Fit. *in Nonparametric Statistical Inference*, pp. 101-148Chapman and Hall.

Gibbons, J.D.a.C., S., 2011b. 3 Tests of Randomness. *in Nonparametric Statistical Inference*, pp. 75-96Chapman and Hall.

Habermann, R.E., 1983. Teleseismic Detection in the Aleutian Island-Arc, *J Geophys Res,* 88**,** 5056-5064.

Hainzl, S. & Ogata, Y., 2005. Detecting fluid signals in seismicity data through statistical earthquake modeling, *J Geophys Res-Sol Ea,* 110.

Hainzl, S., Steacy, S. & Marsan, D., 2010. Seismicity models based on Coulomb stress calculations, *Community Online Resource for Statistical Seismicity Analysis*.

Helmstetter, A., 2003. Is earthquake triggering driven by small earthquakes?, *Phys Rev Lett,* 91.

Jasa, T. & Xiang, N., 2012. Nested sampling applied in Bayesian room-acoustics decay analysis, *J Acoust Soc Am,* 132**,** 3251-3262.

Kass, R.E. & Raftery, A.E., 1995. Bayes Factors, *J Am Stat Assoc,* 90**,** 773-795.

Katsumata, K., 2011. Precursory seismic quiescence before the M-w=8.3 Tokachi-oki, Japan, earthquake on 26 September 2003 revealed by a re-examined earthquake catalog, *J Geophys Res-Sol Ea,* 116.

Llenos, A.L. & Michael, A.J., 2013. Modeling Earthquake Rate Changes in Oklahoma and Arkansas: Possible Signatures of Induced Seismicity, *B Seismol Soc Am,* 103**,** 2850-2861.

Lombardi, A.M., Cocco, M. & Marzocchi, W., 2010. On the Increase of Background Seismicity Rate during the 1997-1998 Umbria-Marche, Central Italy, Sequence: Apparent Variation or Fluid-Driven Triggering?, *B Seismol Soc Am,* 100**,** 1138-1152.

Lombardi, A.M., Marzocchi, W. & Selva, J., 2006. Exploring the evolution of a volcanic seismic swarm: The case of the 2000 Izu Islands swarm, *Geophys Res Lett,* 33.
Ma, L., 2001. Relative quiescence within the Jiashi swarm in Xinjiang, China: an application of the ETAS point process model, *Journal of Applied Probability,* 38**,** 213-221.
Marsan, D. & Nalbant, S.S., 2005. Methods for measuring seismicity rate changes: A review and a study of how the M-w 7.3 Landers earthquake affected the aftershock sequence of the M-w 6.1 Joshua Tree earthquake, *Pure Appl Geophys,* 162**,** 1151-1185.
Matsu'ura, R.S. & Karakama, I., 2005. A point-process analysis of the Matsushiro earthquake swarm sequence: The effect of water on earthquake occurrence, *Pure Appl Geophys,* 162**,** 1319-1345.
Matthews, M.V. & Reasenberg, P.A., 1987. Comment on Habermann's method for detecting seismicity rate changes.
Matthews, M.V. & Reasenberg, P.A., 1988. Statistical-Methods for Investigating Quiescence and Other Temporal Seismicity Patterns, *Pure Appl Geophys,* 126**,** 357-372.
Michael, A.J., 2011. Random variability explains apparent global clustering of large earthquakes, *Geophys Res Lett,* 38.
Michael, A.J., 2014. How complete is the ISC-GEM global earthquake catalog?, *B Seismol Soc Am,* 104**,** 1829-1837.
Mulargia, F. & Tinti, S., 1985. Seismic Sample Areas Defined from Incomplete Catalogs - an Application to the Italian Territory, *Phys Earth Planet In,* 40**,** 273-300.
Naus, J.I., 1982. Approximations for Distributions of Scan Statistics, *J Am Stat Assoc,* 77**,** 177-183.
Naylor, M., Main, I.G., & Touati, S., 2009. Quantifying uncertainty on mean earthquake inter-event times for a finite sample, *J Geophys Res*, 114, B01316.
Ogata, Y., 1988. Statistical-Models for Earthquake Occurrences and Residual Analysis for Point-Processes, *J Am Stat Assoc,* 83**,** 9-27.
Ogata, Y., 1992. Detection of Precursory Relative Quiescence before Great Earthquakes through a Statistical-Model, *J Geophys Res-Sol Ea,* 97**,** 19845-19871.
Ogata, Y., 1998. Space-time point-process models for earthquake occurrences, *Ann I Stat Math,* 50**,** 379-402.
Ogata, Y., 1999. Seismicity analysis through point-process modeling: A review, *Pure Appl Geophys,* 155**,** 471-507.
Ogata, Y., 2007. Seismicity and geodetic anomalies in a wide area preceding the Niigata-Ken-Chuetsu earthquake of 23 October 2004, central Japan, *J Geophys Res-Sol Ea,* 112.
Saichev, A. & Sornette, D., 2006. Renormalization of branching models of triggered seismicity from total to observable seismicity, *Eur Phys J B,* 51**,** 443-459.
Sambridge, M., Gallagher, K., Jackson, A. & Rickwood, P., 2006. Trans-dimensional inverse problems, model comparison and the evidence, *Geophys J Int,* 167**,** 528-542.
Schwarz, G., 1978. Estimating Dimension of a Model, *Ann Stat,* 6**,** 461-464.
Shaw, J.R., Bridges, M. & Hobson, M.P., 2007. Nested sampling for general bayesian computation, *Mon Not R Astron Soc,* 378**,** 1365-1370.
Shearer, P.M. & Stark, P.B., 2012. Global risk of big earthquakes has not recently increased, *P Natl Acad Sci USA,* 109**,** 717-721.
Sivia, D., and Skilling, J., 2007. Nested Sampling Main Program (in R)http://www.inference.phy.cam.ac.uk/bayesys/r/mininest.r.
Skilling, J., 2006. Nested Sampling for General Bayesian Computation, *Bayesian Anal,* 1**,** 833-859.
Storchak, D.A., Di Giacomo, D., Bondar, I., Engdahl, E.R., Harris, J., Lee, W.H.K., Villasenor, A. & Bormann, P., 2013. Public Release of the ISC-GEM Global Instrumental Earthquake Catalogue (1900-2009), *Seismol Res Lett,* 84**,** 810-815.
Touati, S., Naylor, M. & Main, I.G., 2009. Origin and Nonuniversality of the Earthquake Interevent Time Distribution, *Phys Rev Lett,* 102.
Touati, S., Naylor, M. & Main, I.G., 2014. Statistical Modeling of the 1997-1998 Colfiorito Earthquake Sequence: Locating a Stationary Solution within Parameter Uncertainty, *B Seismol Soc Am,* 104**,** 885-897.

Touati, S., Naylor, M., Main, I.G. & Christie, M., 2011. Masking of earthquake triggering behavior by a high background rate and implications for epidemic-type aftershock sequence inversions, *J Geophys Res-Sol Ea,* 116.
Wang, Q., Jackson, D.D. & Zhuang, J.C., 2010. Are spontaneous earthquakes stationary in California?, *J Geophys Res-Sol Ea,* 115.
Wasserman, L., 2000. Bayesian model selection and model averaging, *J Math Psychol,* 44**,** 92-107.
Wyss, M. & Wiemer, S., 2000. Change in the probability for earthquakes in southern California due to the Landers magnitude 7.3 earthquake, *Science,* 290**,** 1334-1338.
Zhuang, J., Ogata, Y. & Vere-Jones, D., 2002. Stochastic declustering of space-time earthquake occurrences, *J Am Stat Assoc,* 97**,** 369-380.
Zhuang, J.C., 2000. Statistical modelling of seismicity patterns before and after the 1990 Oct 5 Cape Palliser earthquake, New Zealand, *New Zeal J Geol Geop,* 43**,** 447-460.
Zhuang, J.C., Chang, C.P., Ogata, Y. & Chen, Y.I., 2005. A study on the background and clustering seismicity in the Taiwan region by using point process models, *J Geophys Res-Sol Ea,* 110.

Figure captions

Figure 1: Probability of observing an average rate of 1.1, 1.5 or 2 over the specified period, under a null hypothesis of a Poisson process with unit rate, based on both the Z-score method that regards the error in the mean event rate to be approximately normally distributed (red dashed line) and the basic Poisson probability (black solid line; colour online).

Figure 2: Summary of the Poisson rate change simulations. (a) Fraction of Poisson simulations for which the null hypothesis of a stationary Poisson process was rejected, for each statistical test, as a function of *v*, the factor by which the simulated rate increases at the changepoint half-way through the simulation. Each point is based on 500 realisations of 1000 events, as detailed in the text. Note that the simple Z test has two curves; for the upper one, the long-term rate was estimated from the first 500 events, and for the lower, the long-term rate was estimated from the whole 1000 events; (b) Cartoon figure to help with the interpretation of parts (a, c); (c) Comparison of the performance of the different methods in terms of the rate of false positives and the threshold at which true rate changes are accepted 95% of the time for the Poisson rate change synthetics. The Z-test methods are linked by a yellow line; the Bayesian methods are shown in a green oval and the arrows are to indicate that the 95% acceptance rate is significantly off to the right of this figure for the runs and K-S tests.

Figure 3: Fraction of ETAS simulations for which the null hypothesis of a stationary Poisson process was rejected, for the Habermann Z test and AIC, as a function of ν, and using different values of μ1. Again, each point is based on 500 realisations of 1000 events.

Figure 4: Histograms of Habermann's Z for the residual point process derived from stationary ETAS simulations, along with the expectation from the standard normal null distribution of Z (in grey, with error bars showing 95% confidence intervals based on Poisson counting errors). For simulations with high μ1, the histogram is significantly different from the expectation.

Figure 5: Probability of a hypothetical cluster of 70 events in 50 days, given a Poisson process with an underlying rate of 1 event per day, and given a varying length of event history in which this cluster is situated. The Poisson scan statistics method takes into account the increased likelihood of a given cluster occurring somewhere in a longer catalogue, while Habermann's Z test does not.

Figure 6: Event rate (per year) averaged over 5-year windows through the global catalogue, for different threshold magnitudes and for all data (top) vs declustered data (bottom).

Tables

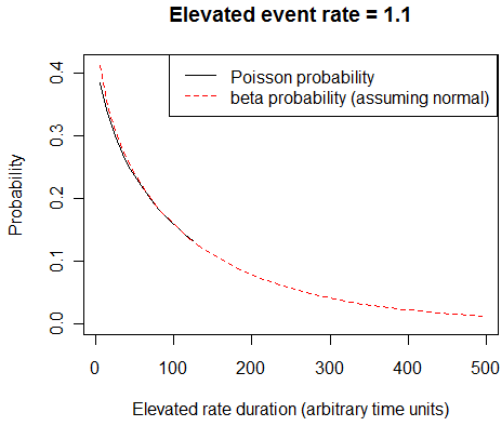
Simulation case	Stationary ETAS model: log evidence (standard deviation)	Changepoint ETAS model: log evidence (standard deviation)	log Bayes Factor	Correct model chosen?
$\mu_1 = 0.01$, $\nu = 1$	-1397.28 (0.49)	-1398.11 (0.56)	-0.83	Y
$\mu_1 = 0.01$, $\nu = 2$	-1388.43 (0.49)	-1383.58 (0.56)	4.85	Y
$\mu_1 = 0.1$, $\nu = 1$	-421.95 (0.45)	-419.53 (0.52)	2.42	N
$\mu_1 = 0.1$, $\nu = 2$	-245.76 (0.45)	-241.28 (0.52)	4.48	Y
$\mu_1 = 1$, $\nu = 1$	961.08 (0.42)	969.10 (0.46)	8.02	N
$\mu_1 = 1$, $\nu = 2$	1286.40 (0.41)	1290.75 (0.46)	4.35	Y

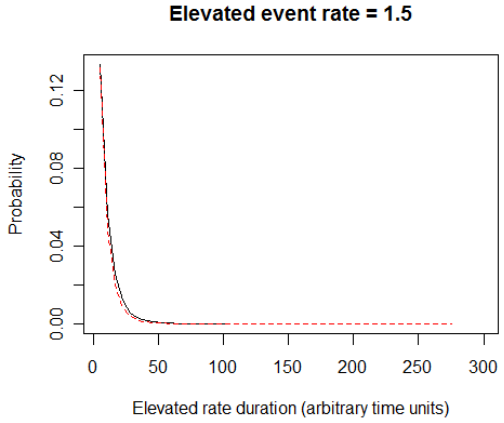
Table 1: The (log) Bayesian Evidence, computed by nested sampling, for the ETAS model using test cases of ETAS simulations. The (log) Bayes Factor obtained from comparing the stationary and changepoint models, with a positive log Bayes Factor indicating a preference for the changepoint model.

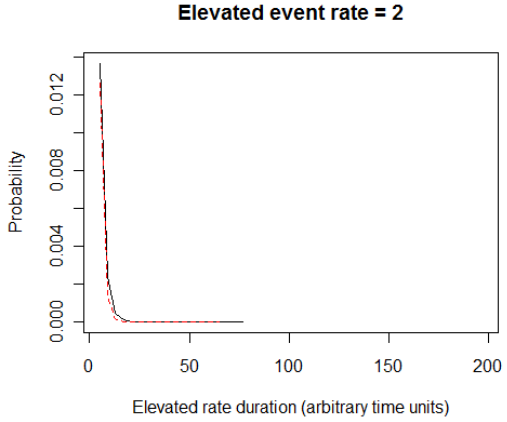
Threshold magnitude	Number of events	Stationary Poisson model parameter (events/yr)	Changepoint Poisson model parameters (events/yr, events/yr, year)	AIC and BIC results	Bayes Factor result	Habermann Z result	Probability of post-changepoint data using Poisson scan statistics
7.0	1114	$\mu = 11.51$	$\mu_1 = 10.73$, $\mu_2 = 14.60$, $t_c = 1995.27$	ΔAIC = 11.10, ΔBIC = 5.07	log BF = -1.60	Z = 4.11	N/A
	935 declustered	$\mu = 9.66$	$\mu_1 = 9.14$, $\mu_2 = 11.29$, $t_c = 1992.32$	ΔAIC = 0.31, ΔBIC = -5.38	log BF = -0.57	Z = 2.77	N/A
7.5	358	$\mu = 3.70$	$\mu_1 = 3.40$, $\mu_2 = 5.36$, $t_c = 2000.24$	ΔAIC = 3.51, ΔBIC = -0.25	log BF = 1.25	Z = 3.06	0.048
	316 declustered	$\mu = 3.26$	$\mu_1 = 3.01$, $\mu_2 = 4.27$, $t_c = 1995.37$	ΔAIC = -0.91, ΔBIC = -4.42	log BF = 0.06	Z = 2.48	0.174
8.0	76	$\mu = 0.79$	$\mu_1 = 0.70$, $\mu_2 = 1.53$, $t_c = 2004.98$	ΔAIC = -1.79, ΔBIC = -2.45	log BF = -1.53	Z = 2.04	0.361
	70 declustered	$\mu = 0.72$	$\mu_1 = 0.67$, $\mu_2 = 1.22$, $t_c = 2004.98$	ΔAIC = -4.80, ΔBIC = -5.30	log BF = -2.16	Z = 1.53	0.771
8.5	16	$\mu = 0.17$	$\mu_1 = 0.13$, $\mu_2 = 0.51$, $t_c = 2004.99$	ΔAIC = -2.63, ΔBIC = -0.17	log BF = -1.93	Z = 1.66	0.439
	14 declustered	$\mu = 0.14$	$\mu_1 = 0.12$, $\mu_2 = 0.37$, $t_c = 2004.00$	ΔAIC = -4.84, ΔBIC = -2.11	log BF = -2.34	Z = 1.35	0.708

Table 2: Results of changepoint detection methods applied to global data of different lower magnitude thresholds and either declustered or not.

1 **Figures**


Elevated event rate = 1.1
Poisson probability
beta probability (assuming normal)
Probability
0.0
0.1
0.2
0.3
0.4
0
100
200
300
400
500
Elevated rate duration (arbitrary time units)


Elevated event rate = 1.5
Probability
0.00
0.04
0.08
0.12
0
50
100
150
200
250
300
Elevated rate duration (arbitrary time units)


Elevated event rate = 2
Probability
0.000
0.004
0.008
0.012
0
50
100
150
200
Elevated rate duration (arbitrary time units)

1

Figure 1: Probability of observing an average rate of 1.1, 1.5 or 2 over the specified period, under a null hypothesis of a Poisson process with unit rate, based on both the Z-score method that regards the error in the mean event rate to be approximately normally distributed (red dashed line) and the basic Poisson probability (black solid line; colour online).

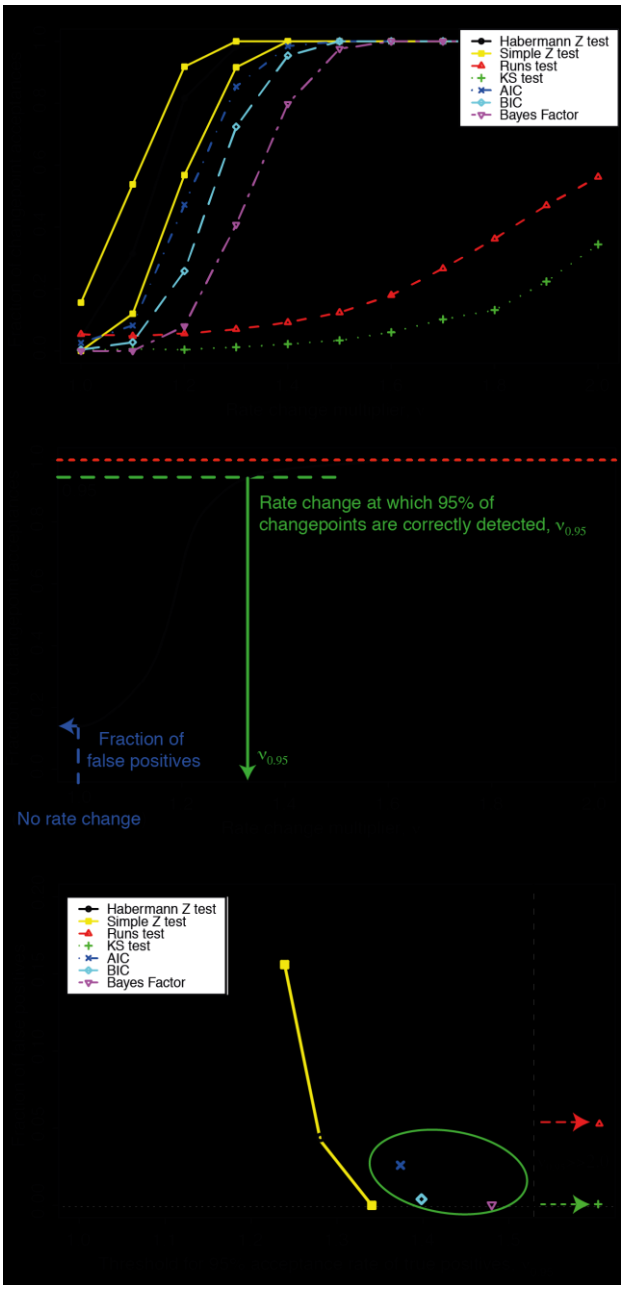


Figure 2: Summary of the Poisson rate change simulations. (a) Fraction of Poisson simulations for which the null hypothesis of a stationary Poisson process was rejected, for each statistical test, as a function of *ν*, the factor by which

the simulated rate increases at the changepoint half-way through the simulation. Each point is based on 500 realisations of 1000 events, as detailed in the text. Note that the simple Z test has two curves; for the upper one, the long-term rate was estimated from the first 500 events, and for the lower, the long-term rate was estimated from the whole 1000 events; (b) Cartoon figure to help with the interpretation of parts (a, c); (c) Comparison of the performance of the different methods in terms of the rate of false positives and the threshold at which true rate changes are accepted 95% of the time for the Poisson rate change synthetics. The Z-test methods are linked by a yellow line; the Bayesian methods are shown in a green oval and the arrows are to indicate that the 95% acceptance rate is significantly off to the right of this figure for the runs and K-S tests.

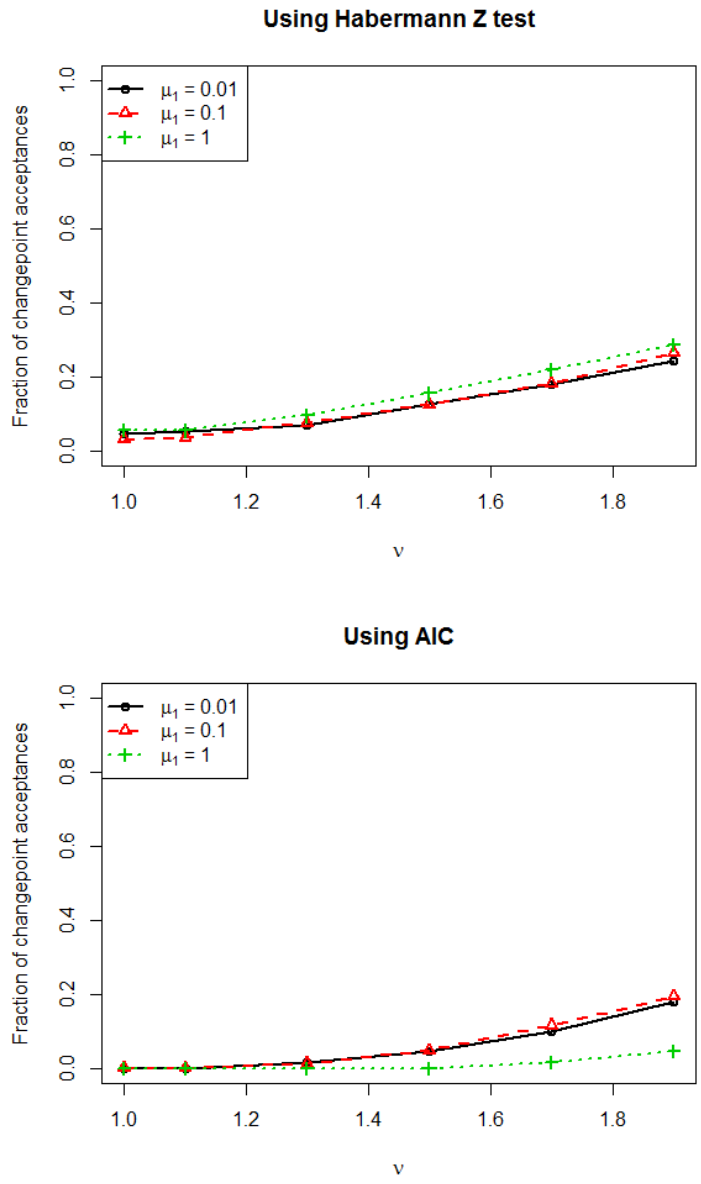


Figure 3: Fraction of ETAS simulations for which the null hypothesis of a stationary Poisson process was rejected, for the Habermann Z test and AIC, as a function of ν, and using different values of μ_1. Again, each point is based on 500 realisations of 1000 events.

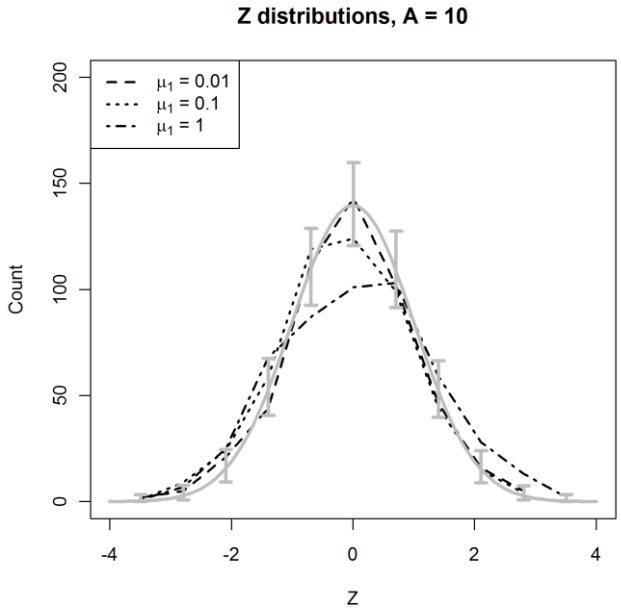


Figure 4: Histograms of Habermann's Z for the residual point process derived from stationary ETAS simulations, along with the expectation from the standard normal null distribution of Z (in grey, with error bars showing 95% confidence intervals based on Poisson counting errors). For simulations with high μ_1, the histogram is significantly different from the expectation.

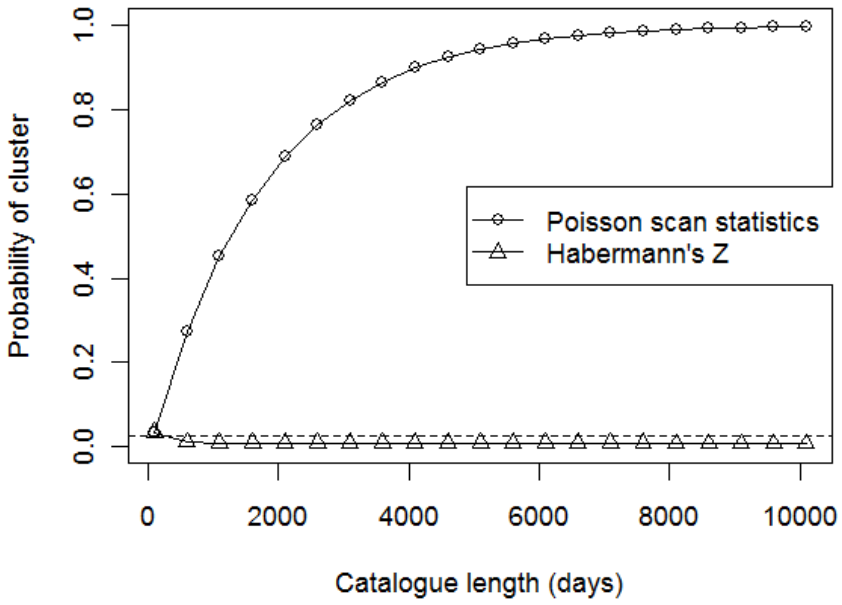


Figure 5: Probability of a hypothetical cluster of 70 events in 50 days, given a Poisson process with an underlying rate of 1 event per day, and given a varying length of event history in which this cluster is situated. The Poisson scan statistics method takes into account the increased likelihood of a given cluster occurring somewhere in a longer catalogue, while Habermann's Z test does not.

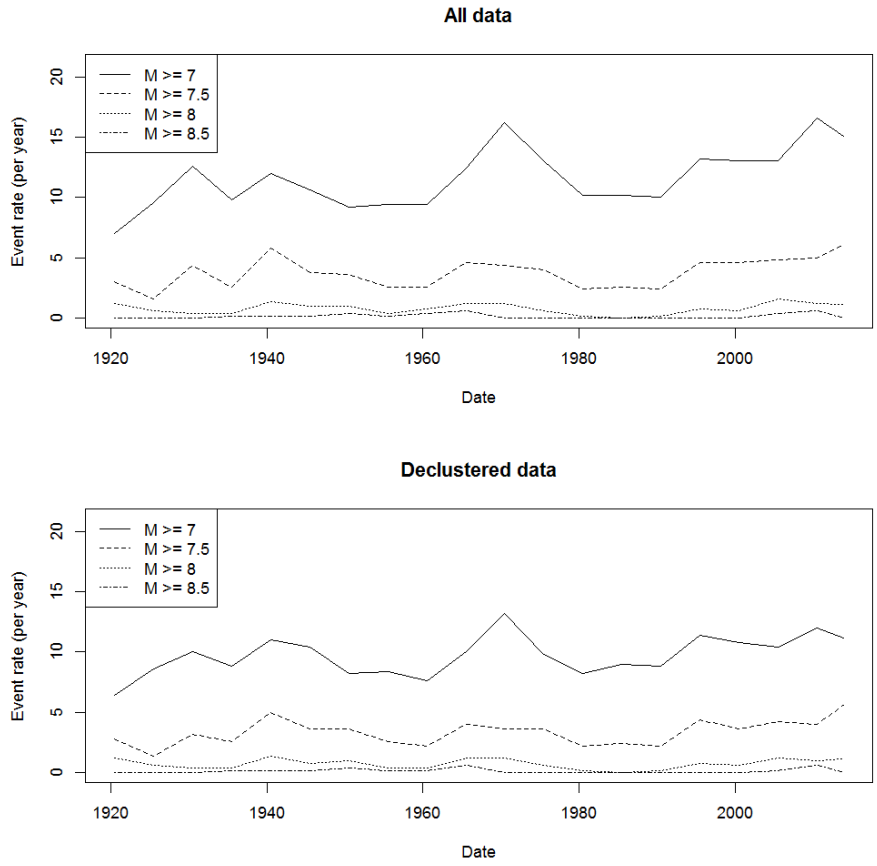


Figure 6: Event rate (per year) averaged over 5-year windows through the global catalogue, for different threshold magnitudes and for all data (top) vs declustered data (bottom).